Mach Learn (2017) 106:387–418
DOI 10.1007/s10994-016-5621-5

Hierarchical Dirichlet scaling process

Dongwoo Kim[1] · Alice Oh[2]

Received: 7 December 2015 / Accepted: 30 November 2016 / Published online: 26 January 2017

Abstract We present the *hierarchical Dirichlet scaling process* (HDSP), a Bayesian nonparametric mixed membership model. The HDSP generalizes the hierarchical Dirichlet process to model the correlation structure between metadata in the corpus and mixture components. We construct the HDSP based on the normalized gamma representation of the Dirichlet process, and this construction allows incorporating a scaling function that controls the membership probabilities of the mixture components. We develop two scaling methods to demonstrate that different modeling assumptions can be expressed in the HDSP. We also derive the corresponding approximate posterior inference algorithms using variational Bayes. Through experiments on datasets of newswire, medical journal articles, conference proceedings, and product reviews, we show that the HDSP results in a better predictive performance than labeled LDA, partially labeled LDA, and author topic model and a better negative review classification performance than the supervised topic model and SVM.

Keywords Topic modeling · Dirichlet process · Hierarchical Dirichlet process

1 Introduction

The hierarchical Dirichlet process (HDP) is an important nonparametric Bayesian prior for mixed membership models, and the HDP topic model is useful for a wide variety of tasks involving unstructured text (Teh et al. 2006). To extend the HDP topic model, there has been active research in dependent random probability measures as priors for modeling the

Editor: Prasad Tadepalli.

✉ Alice Oh
alice.oh@kaist.edu

Dongwoo Kim
dongwoo.kim@anu.edu.au

1 The Australian National University, Acton, Australia

2 KAIST, Daejeon, Korea

underlying association between the latent semantic structure and covariates, such as time stamps and spatial coordinates (Ahmed and Xing 2010; Ren et al. 2011).

A large body of this research is rooted in the dependent Dirichlet process (DDP) (MacEachern 1999) where the probabilistic random measure is defined as a function of covariates. Most DDP approaches rely on the generalization of Sethuraman's stick breaking representation of DP (Sethuraman 1991), incorporating the time difference between two or more data points, the spatial difference among observed data, or the ordering of the data points into the predictor dependent stick breaking process (Duan et al. 2007; Dunson and Park 2008; Griffin and Steel 2006). Some of these priors can be integrated into the hierarchical construction of DP (Srebro and Roweis 2005), resulting in topic models where temporally- or spatially-proximate data are more likely to be clustered. These existing DP approaches, however, cannot be easily extended to model underlying topics of a document collection. One reason is that the extension requires to develop a new tractable inference algorithm from models with intractable posterior distributions.

We suggest the *hierarchical Dirichlet scaling process* (HDSP) as a new way of modeling a corpus with various types of covariates such as categories, authors, and numerical ratings. The HDSP models the relationship between topics and covariates by generating dependent random measures in a hierarchy, where the first level is a Dirichlet process, and the second level is a *Dirichlet scaling process* (DSP). The first level DP is constructed in the traditional way of a stick breaking process, and the second level DSP with a normalized gamma process. With the normalized gamma process, each topic proportion of a document is independently drawn from a gamma distribution and then normalized. Unlike the stick breaking process, the normalized gamma process keeps the same order of the atoms as the first level measure, which allows the topic proportions in the random measure to be controlled. The DSP then uses that controllability to guide the topic proportions of a document by replacing the rate parameter of the gamma distribution with a scaling function that defines the correlation structure between topics and labels. The choice of the scaling function reflects the characteristics of the corpus. We show two scaling functions, the first one for a corpus with categorical labels, and the second for a corpus with both categorical and numerical labels.

The HDSP models the topic proportions of a document as a dependent variable of observable side information. This modeling approach differs from the traditional definition of a generative process where the observable variables are generated from a latent variable or parameter. For example, Zhu et al. (2009) and Mcauliffe and Blei (2007) propose generative processes where the observable labels are generated from a topic proportion of a document. However, a more natural model of the human writing process is to decide what to write about (e.g., categories) before writing the content of a document. This same approach is also successfully demonstrated in Mimno and McCallum (2012).

The outline of this paper is as follows. In Sect. 2, we describe related work and position our work within the topic modeling literature. In Sect. 3, we describe the gamma process construction of the HDP and how scale parameters are used to develop the HDSP with two different scaling functions. In Sect. 4, we derive a variational inference for the latent variables. In Sect. 5, we verify our approach on a synthetic dataset and demonstrate the improved predictive power on real world corpora. In Sect. 6, we discuss our conclusions and possible directions for future work.

2 Related work

For model construction, the model most closely related to HDSP is the discrete infinite logistic normal (DILN) model (Paisley et al. 2012) in which the correlations among topics are modeled through the normalized gamma construction. DILN allocates a latent location for each topic in the first level, and then draws the second level random measures from the normalized gamma construction of the DP. Those random measures are then scaled by an exponentiated Gaussian process defined on the latent locations. DILN is a nonparametric counterpart of the correlated topic model (Blei and Lafferty 2007) in which the logistic normal prior is used to model the correlations between topics. The HDSP is also constructed through the normalized gamma distribution with an informative scaling parameter, but our goal in HDSP is to model the correlations between topics and labels. The doubly correlated nonparametric topic model (DCNT) proposed by Kim and Sudderth (2011) also takes documents' metadata into account to model the correlation among topics and metadata. Unlike the HDSP, the DCNT is constructed through a logistic stick-breaking process (Ren et al. 2011) which is originally proposed for modeling contiguous and spatially localized segments.

The Dirichlet-multinomial regression topic model (DMR-TM) (Mimno and McCallum 2012) also models the label dependent topic proportions of documents, but it is a parametric model. The DMR-TM places a log-linear prior on the parameter of the Dirichlet distribution to incorporate arbitrary types of observed labels. The DMR-TM takes the "upstream" approach in which the latent variable or latent topics are conditionally generated from the observed label information. The author-topic model (Rosen-Zvi et al. 2004) also takes the same approach, but it is a specialized model for authors of documents. Unlike the "downstream" generative approach used in the supervised topic model (Mcauliffe and Blei 2007), the maximum margin topic model (Zhu et al. 2009), and the relational topic model (Chang and Blei 2009), the upstream approach does not require specifying the probability distribution over all possible values of observed labels.

The HDSP is a new way of constructing a dependent random measure in a hierarchy. In the field of Bayesian nonparametrics, the introduction of DDP (Sethuraman 1991) has led to increased attention in constructing dependent random measures. Most such approaches develop priors to allow covariate dependent variation in the atoms of the random measure (Gelfand et al. 2005; Rao and Teh 2009) or in the weights of atoms (Griffin and Steel 2006; Duan et al. 2007; Dunson and Park 2008). These priors replace the first level of the HDP to incorporate a document-specific covariate for generating a dependent topic proportion. The HDSP allows covariate dependent variation in the weights of atoms, where the variation is controlled by the scaling function that defines the correlation between atoms and labels. A proper definition of the scaling function gives the flexibility to model various types of labels.

Several topic models for labeled documents use the credit attribution approach where each observed word token is assigned to one of the observed labels. Labeled LDA (L-LDA) allocates one dimension of the topic simplex per label and generates words from only the topics that correspond to the labels in each document (Ramage et al. 2009). An extension of this model, partially labeled LDA (PLDA), adds more flexibility by allocating a pre-defined number of topics per label and including a background label to handle documents with no labels (Ramage et al. 2011). The Dirichlet process with mixed random measures (DP-MRM) is a nonparametric topic model which generates an unbounded number of topics per label but still excludes topics from labels that are not observed in the document (Kim et al. 2012).

3 Hierarchical Dirichlet scaling process

In this section, we describe the hierarchical Dirichlet scaling process (HDSP). First we review the HDP with an alternative construction using the normalized gamma process construction for the second level DP. We then present the HDSP where the second level DP is replaced by Dirichlet scaling process (DSP). Finally, we describe two scaling functions for the DSP to incorporate categorical and numerical labels.

3.1 The normalized gamma process construction of HDP

The HDP[1] consists of two levels of the DP where the random measure drawn from the upper level DP is the base distribution of the lower level DP. The formal definition of the hierarchical representation is as follows:

$$G_0 \sim \mathrm{DP}(\alpha, H), \qquad G_m \sim \mathrm{DP}(\beta, G_0), \tag{1}$$

where H is a base distribution, α, and β are concentration parameters for each level respectively, and index m represents multiple draws from the second level DP. For the mixed membership model, x_{mn}, observation n in group m, can be drawn from

$$\theta_{mn} \sim G_m, \qquad \mathrm{x}_{mn} \sim f(\theta_{mn}), \tag{2}$$

where $f(\cdot)$ is a data distribution parameterized by θ. In the context of topic models, the base distribution H is usually a Dirichlet distribution over the vocabulary, so the atoms of the first level random measure G_0 are an infinite set of topics drawn from H. The second level random measure G_m is distributed based on the first level random measure G_0, so the second level shares the same set of topics, the atoms of the first level random measure.

The constructive definition of the DP can be represented as a stick breaking process (Sethuraman 1991), and in the HDP inference algorithm based on stick breaking, the first level DP is given by the following conditional distributions:

$$\begin{aligned} V_k &\sim \mathrm{Beta}(1, \alpha) & p_k &= V_k \prod_{j=1}^{j<k} (1 - V_j) \\ \phi_k &\sim H & G_0 &= \sum_{k=1}^{\infty} p_k \delta_{\phi_k}, \end{aligned} \tag{3}$$

where V_k defines a corpus level topic distribution for topic ϕ_k. The second level random measures are conditionally distributed on the first level discrete random measure G_0:

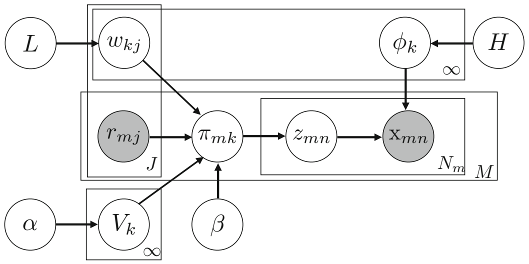
$$\begin{aligned} \pi_{ml} &\sim \mathrm{Beta}(1, \beta) & p_{ml} &= \pi_{ml} \prod_{j=1}^{j<l} (1 - \pi_{mj}) \\ \theta_{ml} &\sim G_0 & G_m &= \sum_{l=1}^{\infty} p_{ml} \delta_{\theta_{ml}}, \end{aligned} \tag{4}$$

where the second level atom θ_{ml} corresponds to one of the first level atoms ϕ_k.

An alternative construction of the HDP is based on the normalized gamma process (Paisley et al. 2012). While the first level construction remains the same, the gamma process changes the second level construction from Eq. 4 to

$$\begin{aligned} \pi_{mk} &\sim \mathrm{Gamma}(\beta p_k, 1) \\ G_m &= \sum_{k=1}^{\infty} \frac{\pi_{mk}}{\sum_{j=1}^{\infty} \pi_{mj}} \delta\phi_k, \end{aligned} \tag{5}$$

[1] In this paper, we limit our discussions of the HDP to the two level construction of the DP and refer to it simply as the HDP.



Notations

α, β	: concentration parameters
H, L	: base distribution for topics
w_{kj}	: weight between topic *k* and label *j*
ϕ_k	: *k* th topic
V_k, π_{mk}	: first and second level topic proportion
z_{mn}	: topic of *n* th word in *m* th document
r_{mj}	: *j* th label of *m* th document {0, 1}
x_{mn}	: observed *n* th word in *m* th document

Fig. 1 Graphical model of the hierarchical Dirichlet scaling process

where $\text{Gamma}(x; a, b) = b^a x^{(a-1)} e^{-bx} / \Gamma(a)$. Unlike the stick breaking construction, the atom of the π_{mk} of the gamma process is the same as the atom of the kth stick of the first level. Therefore, during inference, the model does not need to keep track of which second level atoms correspond to which first level atoms. Furthermore, by placing a proper random variable on the rate parameter of the gamma distribution, the model can infer the correlations among the topics (Paisley et al. 2012) through the Gaussian process (Rasmussen and Williams 2005).

The normalized gamma process itself is not an appropriate construction method for the approximate posterior inference algorithm based on the variational truncation method (Blei and Jordan 2006) because, unlike the stick breaking process, the probability mass of a random measure constructed by the normalized gamma process is not limited to the first few number of atoms. But once the base distribution of second level DP is constructed by the stick breaking process of first level DP, the total mass of the second level base distribution G_0 is limited to the first few number of atoms, and then the truncation based posterior inference algorithm approximates the true posterior of the normalized gamma construction (Fig. 1).

3.2 Hierarchical Dirichlet scaling process

The HDSP generalizes the HDP by modeling mixture proportions dependent on covariates. As a topic model, the HDSP assumes that topics and labels are correlated, and the topic proportions of a document are proportional to the correlations between the topics and the observed labels of the document. We develop the Dirichlet scaling process (DSP) with the normalized gamma construction of the DP, where the rate parameter of the gamma distribution is replaced by the scaling function. This scaling function serves the central role of defining the correlation structure between a topic and labels. Formally, the HDSP consists of DP and DSP in a hierarchy:

$$G_0 \sim \text{DP}(\alpha, H) \tag{6}$$

$$G_m \sim \text{DSP}(\beta, G_0, r_m, s_w(\cdot)), \tag{7}$$

where the first level random measure G_0 is drawn from the DP with concentration parameter α and base distribution H. The second level random measure G_m for document m is drawn from the DSP parameterized by the concentration parameter β, base distribution G_0, observed labels of document r_m, and *scaling function* $s(\cdot)$ with *scaling parameter* w.

As in the HDP, the first level of HDSP is a DP where the base distribution is the product of two distributions for data distribution and scaling parameter w. Specifically, the base distribution H is $\text{Dir}(\eta) \otimes L_w$ where η is the parameter of the word-topic distribution, and

L_w is a prior distribution for the scaling parameter w. The form of the resulting random measure is

$$G_0 = \sum_{k=1}^{\infty} p_k \delta_{\{\phi_k, w_k\}}, \tag{8}$$

where p_k is the stick length for topic k, $p_k = V_k \prod_{k'=1}^{k'<k}(1 - V_{k'})$ and $\{\phi_k, w_k\}$ is the atom of stick k. At the second level construction, w_k becomes the parameter to guide the proportion of topic k's for each document.

At the second level of HDSP, label-dependent random measures are drawn from the DSP. First, as in the HDP, draw a random measure $G'_m \sim \mathrm{DP}(\beta, H)$ for document m. Second, scale the weights of the atoms based on a scaling function parameterized by w_k and the observed labels. Let r_{mj} be the value of observed label j in document m, then G'_m is scaled as follows:

$$G_m(\{\phi_k, l_k\}) \propto G'_m(\{\phi_k, l_k\}) \times s_{w_k}(r_{mj}) \tag{9}$$

where $s_{w_k}(\cdot)$ is the scaling function parameterized by the scaling parameter w_k. Topic k is scaled by the scaling weight, $s_{w_k}(r_{mj})$, and therefore, the topic proportions of a document is proportional to the scaling weights of the observed labels. The scaling function should be carefully chosen to reflect the underlying relationship between topics and labels. We show two concrete examples of scaling functions in Sect. 3.3.

The constructive definition of HDSP is similar to the HDP, but the difference comes from the scaling function. The stick breaking process is used to construct the first level random measure:

$$\begin{aligned} &V_k \sim \mathrm{Beta}(1, \alpha) && p_k = V_k \prod_{j=1}^{j<k}(1 - V_j) \\ &\phi_k \sim \mathrm{Dir}(\eta), \quad w_k \sim L_w && G_0 = \sum_{k=1}^{\infty} p_k \delta_{\{\phi_k, w_k\}}, \end{aligned} \tag{10}$$

where the pair $\{\phi_k, w_k\}$ drawn i.i.d. from two base distributions forms an atom of the resulting measure.

Based on the discrete first level random measure, the second level random measure is constructed by the normalized gamma process. As in the HDP, the weight of atom k is drawn from a gamma distribution with parameter βp_k, and then scaled by the scaling weight $s_{w_k}(r_m)$

$$\pi_{mk} \sim \mathrm{Gamma}(\beta p_k, 1) \times s_{w_k}(r_m). \tag{11}$$

The scaling weight can be directly incorporated into the second parameter of the gamma distribution because the scaled gamma random variable $y = kx \sim \mathrm{Gamma}(a, 1)$ is equal to $y \sim \mathrm{Gamma}(a, k^{-1})$,

$$\pi_{mk} \sim \mathrm{Gamma}(\beta p_k, s_{w_k}(r_m)^{-1}). \tag{12}$$

Then, the random variables are normalized to form a proper probability random measure

$$G_m = \sum_{k=1}^{\infty} \frac{\pi_{mk}}{\sum_{j=1}^{\infty} \pi_{mj}} \delta_{\phi_k}. \tag{13}$$

For the mixed membership model, nth observation in mth group is drawn as follows:

$$\phi_k \sim G_m, \quad \mathrm{x}_{mn} \sim f(\phi_k), \tag{14}$$

where f is a data distribution parameterized by ϕ_k. For topic modeling, G_m and x_{mn} correspond to document m and word n in document m, respectively.

3.3 Scaling functions

Now we propose two scaling functions to express the correlation between topics and labels of documents. A scaling method is properly defined by two factors: 1) a proper prior over the scaling parameter w_k, 2) a plausible scaling function between topic specific scaling parameter w_k and the observed labels of document r_m.

Scaling function 1 We design the first scaling function to model categorical side information such as authors, tags, and categories. For a corpus with J unique labels, then w_k is a J-dimensional parameter where each dimension matches to a corresponding label. We define the scaling function as the product of scaling parameters that correspond to the observed labels:

$$s_{w_k}(r_m) = \prod_{j=1}^{J} w_{kj}^{r_{mj}} \qquad w_{kj} \sim \text{inv-Gamma}(a_w, b_w) \tag{15}$$

where r_{mj} is an indicator variable whose value is one when label j is observed in document m and zero otherwise. w_{kj} is a scaling parameter of topic k for label j. We place a inverse gamma prior over the weight variable w_{kj}.

With this scaling function, the proportion of topic k for document m is scaled as follows:

$$\pi_{mk} \sim \text{Gamma}(\beta p_k, 1) \times \prod_{j=1}^{J} w_{kj}^{r_{mj}}. \tag{16}$$

The scaled gamma distribution is equal to the gamma distribution with the rate parameter of inverse scaling factor, so we can rewrite the above equation as follows:

$$\pi_{mk} \sim \text{Gamma}\left(\beta p_k, \prod_{j=1}^{J} w_{kj}^{-r_{mj}}\right). \tag{17}$$

Finally, we normalize these random variables to make a probabilistic random measure summed up to unity for document m:

$$\bar{\pi}_{mk} = \frac{\pi_{mk}}{\sum_{k'} \pi_{mk'}}. \tag{18}$$

Scaling function 2 The above scaling function models categorical side information, but many datasets, such as product reviews have numerical ratings as well as categorical information. We propose the second scaling function that can model both numerical and categorical information. Again, let w_k be J-dimensional scaling parameter where each dimension matches to a corresponding label. The second scaling function is defined as follows:

$$s_{w_k}(r_m) = \frac{1}{\exp\left(\sum_j w_{kj} r_{mj}\right)}, \tag{19}$$

where w_{kj} is the scaling parameter of label j for topic k, and r_{mj} is the observed value of label j of document m. We place a normal prior over the scaling parameter w_k. The scaling function is an inverse log-linear to the weighted sum of document's labels. Unlike the previous scaling function which only considers whether a label is observed in a document, this scaling function incorporates the value of the observed label. With this scaling function, the proportion of topic k for document m is scaled as follows

$$\pi_{mk} \sim \text{Gamma}(\beta p_k, 1) \times \frac{1}{\exp\left(\sum_j w_{kj} r_{mj}\right)}. \tag{20}$$

Again, we can rewrite this equations as

$$\pi_{mk} \sim \text{Gamma}\left(\beta p_k, \exp\left(\sum_j w_{kj} r_{mj}\right)\right). \tag{21}$$

π_{mk} is proportional to the inverse weighted sum of observed labels. Again, we normalize π_{mk} to construct a proper random measure.

The choice of scaling function reflects the modeler's perspective with respect to the underlying relationship between topics and labels. The first scaling function scales each topic by the product of the scaling parameters of the observed labels. This reflects the modeler's assumption that a document with a set of observed labels is likely to exhibit topics that have high correlation with all of the observed labels. With the second scaling function, the scaling weight changes exponentially as the value of label changes. This reflects the modeler's assumption that two documents with the same set of observed labels but with different values are likely to exhibit different topics.

3.4 HDSP as a dependent Dirichlet process

We can view the HDSP as an alternative construction of the hierarchical dependent Dirichlet process (DDP) via a hierarchy consisting of a stick breaking process and a normalized gamma process. Let us compare the HDSP approach to the general DDP approach for topic modeling. The formal definition of DDP is:

$$G_0(\cdot) \sim \text{DDP}(\alpha, H), \tag{22}$$

where the resulting random measure G_0 is a function of some covariates. Using G_0 as the base distribution of a DP for a document with a covariate, the random measure corresponding to document m is constructed as follows:

$$G_m \sim \text{DP}(\beta, G_0(r_m)), \tag{23}$$

where $G_0(r_m)$ is the base distribution for the document with same covariate r_m (Srebro and Roweis 2005).

Similarly, the HDSP constructs a dependent random measure with covariates. However, unlike the DDP-DP approach, G_0 is no longer a function of covariates. The HDSP defines a single global random measure G_0 and then scales G_0 based on the covariates with the scaling function. With a proper, but relatively simple, scaling function that reflects the correlation between covariates and topics, the HDSP models any structures or types of covariates, whereas the DDP requires a complex dependent process for different types of covariates (Griffin and Steel 2006).

4 Variational inference for HDSP

The posterior inference for Bayesian nonparametric models is important because it is intractable to compute the posterior over an infinite dimensional space. Approximation algorithms, such as marginalized MCMC (Escobar and West 1995; Teh et al. 2006) and variational inference (Blei and Jordan 2006; Teh et al. 2008), have been developed for the Bayesian nonparametric mixture models. We develop a mean field variational inference (Jordan et al. 1999; Wainwright and Jordan 2008) algorithm for approximate posterior inference of the HDSP topic model. The objective of variational inference is to minimize the KL divergence between a distribution over the hidden variables and the true posterior, which is equivalent to maximizing the lower bound of the marginal log likelihood of observed data.

In this section, we first derive the inference algorithm for the first scaling function with a fully factorized variational family. Variational inference algorithms can be easily modularized with the fully factorized variational family, and the variation in a model only affects the update rules for the modified parts of the model. Therefore, for the second scaling function, we only need to update the part of the inference algorithm related to the new scaling function.

4.1 Variational inference for the first scaling function

For the first scaling function, we use a fully factorized variational distribution and perform a mean-field variational inference. There are five latent variables of interest: the corpus level stick proportion V_k, the document level stick proportion π_{mk}, the scaling parameter between topic and label w_{kj}, the topic assignment for each word z_{mn}, and the word topic distribution ϕ_k. Thus the variational distribution $q(z, \pi, V, w, \phi)$ can be factorized into

$$
\begin{aligned}
&q(z, \pi, V, w, \phi) = \\
&\prod_{k=1}^{T} \prod_{m=1}^{M} \prod_{j=1}^{J} \prod_{n=1}^{N_m} q(z_{mn}) q(\pi_{mk}) q(V_k) q(\phi_k) q(w_{kj}),
\end{aligned}
\tag{24}
$$

where the variational distributions are

$$
\begin{aligned}
q(z_{mn}) &= \text{Multinomial}(z_{mn} | \gamma_{mn}) \\
q(\pi_{mk}) &= \text{Gamma}(\pi_{mk} | a^{\pi}_{mk}, b^{\pi}_{mk}) \\
q(V_k) &= \delta_{V_k} \\
q(\phi_k) &= \text{Dirichlet}(\phi_k | \eta_k) \\
q(w_{kj}) &= \text{InvGamma}(w_{kj} | a^{w}_{kj}, b^{w}_{kj}).
\end{aligned}
$$

For the corpus level stick proportion V_k, we use the delta function as a variational distribution for simplicity and tractability in inference steps as demonstrated in Liang et al. (2007). Infinite dimensions over the posterior is a key problem in Bayesian nonparametric models and requires an approximation method. In variational treatment, we truncate the unbounded dimensionality to T by letting $V_T = 1$. Thus the model still keeps the infinite dimensionality while allowing approximation to be carried out under the bounded variational distributions.

Using standard variational theory, we derive the evidence lower bound (ELBO) of the marginal log likelihood of the observed data $\mathcal{D} = (\mathbf{x}_m, \mathbf{r}_m)_{m=1}^{M}$,

$$
\begin{aligned}
&\log p(\mathcal{D} | \alpha, \beta, a^w, b^w, \eta) \\
&\quad \geq \mathbb{E}_q[\log p(\mathcal{D}, z, \pi, V, w, \phi)] + H(q) = \mathcal{L}(q),
\end{aligned}
\tag{25}
$$

where $H(q)$ is the entropy for the variational distribution. By taking the derivative of this lower bound, we derive the following coordinate ascent algorithm.

Document-level updates At the document level, we update the variational distribution for the topic assignment z_{mn} and the document level stick proportion π_{mk}. The update for $q(z_{mn}|\gamma_{mn})$ is

$$\gamma_{mnk} \propto \exp\left(\mathbb{E}_q[\ln \eta_{k,x_{mn}}] + \mathbb{E}_q[\ln \pi_{mk}]\right). \tag{26}$$

Updating $q(\pi_{mk}|a^{\pi}_{mk}, b^{\pi}_{mk})$ requires computing the expectation term $\mathbb{E}[\ln \sum_{k=1}^{T} \pi_{mk}]$. Following Blei and Lafferty (2007), we approximate the lower bound of the expectation by using the first-order Taylor expansion,

$$-\mathbb{E}_q\left[\ln \sum_{k=1}^{T} \pi_{mk}\right] \geq -\ln \xi_m - \frac{\sum_{k=1}^{T} \mathbb{E}_q[\pi_{mk}] - \xi_m}{\xi_m}, \tag{27}$$

where the update for $\xi_m = \sum_{k=1}^{K} \mathbb{E}_q[\pi_{mk}]$. Then, the update for π_{mk} is

$$\begin{aligned} a^{\pi}_{mk} &= \beta p_k + \sum_{n=1}^{N_m} \gamma_{mnk} \\ b^{\pi}_{mk} &= \prod_j \mathbb{E}_q[w_{kj}^{-r_{mj}}] + \frac{N_m}{\xi_m}. \end{aligned} \tag{28}$$

Note again r_{mj} is equal to 1 when jth label is observed in mth document, otherwise 0.

Corpus-level updates At the corpus level, we update the variational distribution for the scaling parameter w_{kj}, corpus level stick length V_k and word topic distribution η_{ki}.

The optimal form of a variational distribution can be obtained by exponentiating the variational lower bound with all expectations except the parameter of interest (Bishop and Nasrabadi 2006). For w_{kj}, we can derive the optimal form of variational distribution as follows

$$\begin{aligned} q(w_{kj}) &\sim \text{InvGamma}(a', b') \\ a' &= \mathbb{E}_q[\beta p_k] \sum_m r_{mj} + a^w \\ b' &= \sum_{m'} \prod_{j'/j} \mathbb{E}_q[w_{j'k}^{-1}] \mathbb{E}_q[\pi_{m'k}] + b^w, \end{aligned} \tag{29}$$

where $m' = \{m : r_{mj} = 1\}$ and $j'/j = \{j' : r_{mj'} = 1, j' \neq j\}$. See "Appendix 1" for the complete derivation. There is no closed form update for V_k, instead we use the steepest ascent algorithm to jointly optimize V_k. The gradient of V_k is

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial V_k} = &-\frac{\alpha - 1}{1 - V_k} \\ &- \frac{\beta p_k}{V_k} \left\{ \sum_{m,j} r_{mj} \mathbb{E}_q[\ln \pi_{mk}] - \mathbb{E}_q[\pi_{mk}] + \psi(\beta p_k) \right\} \\ &+ \sum_{k'>k} \frac{\beta p_{k'}}{1 - V_k} \left\{ \sum_{m,j} r_{mj} \mathbb{E}_q[\ln \pi_{mk'}] - \mathbb{E}_q[\pi_{mk'}] + \psi(\beta p_{k'}) \right\}, \end{aligned} \tag{30}$$

where $\psi(\cdot)$ is a digamma function. Finally, the update for the word topic distribution $q(\phi_k|\eta_k)$ is

$$\eta_{ki} = \eta + \sum_{m,n} \gamma_{mnk} \mathbf{1}(\mathrm{x}_{mn} = i), \tag{31}$$

where i is a word index, and $\mathbf{1}$ is an indicator function (Blei et al. 2003).

The expectations of latent variables under the variational distribution q are

$$\begin{aligned}
\mathbb{E}_q[\pi_{mk}] &= a^{\pi}_{mk}/b^{\pi}_{mk} \\
\mathbb{E}_q[\ln \pi_{mk}] &= \psi(a^{\pi}_{mk}) - \ln b^{\pi}_{mk} \\
\mathbb{E}_q[w_{kj}] &= b^{w}_{kj}/(a^{w}_{kj} - 1) \\
\mathbb{E}_q[w^{-1}_{kj}] &= a^{w}_{kj}/b^{w}_{kj} \\
\mathbb{E}_q[\ln w_{kj}] &= \ln b^{w}_{kj} - \psi(a^{w}_{kj}) \\
\mathbb{E}_q[\ln \phi_{ki}] &= \psi(\eta_{ki}) - \psi\left(\sum_i \eta_{ki}\right).
\end{aligned}$$

4.2 Variational inference for the second scaling function

Introducing a new scaling function requires a new approximation method. We first choose the part of ELBO which requires new treatment as the scaling function changes. From Eq. 25, we take the terms that are related to the scaling function s:

$$\begin{aligned}
\mathcal{L}_s &= \mathbb{E}_q\left[\sum_{m=1}^{M}\sum_{k=1}^{\infty} \ln p(\pi_{mk}|V_k, s, \mathbf{r}_m)\right] + \mathbb{E}_q[\ln p(s)] - \mathbb{E}_q[\ln q(s)] \\
&= \sum_m \left[\beta p_k \mathbb{E}_q[\ln s(r_m)] + (\beta p_k - 1)\mathbb{E}_q[\ln \pi_{mk}] - \mathbb{E}_q[s(r_m)]\mathbb{E}_q[\pi_{mk}] - \ln \Gamma(\beta p_k)\right] \\
&\quad + \mathbb{E}_q[p(s)] - \mathbb{E}_q[q(s)].
\end{aligned} \tag{32}$$

To update the scaling parameters, we need a proper prior and variational distribution. For the second scaling function, the normal distribution with zero mean and variance σ is used as a prior of w_{kj}, and the delta function is used as the variational distribution of w_{kj}. Newton-Raphson optimization method are used to update the weight parameters. The Newton-Raphson optimization finds a stationary point of a function by iterating:

$$w_k^{\text{new}} \leftarrow w_k^{\text{old}} - H(w_k)^{-1} \frac{\partial \mathcal{L}}{\partial w_k}, \tag{33}$$

where $H(w_k)$ and $\frac{\partial \ell}{\partial w_k}$ are the Hessian matrix and gradient at the point w_k^{old}. The lower bound with respect to the parameter w_{kj} is,

$$\mathcal{L}_{w_{kj}} = \sum_m \left[\beta p_k \sum_j w_{kj} r_{mj} + (\beta p_k - 1)\mathbb{E}_q[\ln \pi_{mk}\right] - \exp(w_k^{\top} r_m)\mathbb{E}_q[\pi_{mk}] - \ln \Gamma(\beta p_k)]. \tag{34}$$

Then, the gradient and Hessian matrix of w_{kj} are

$$\frac{\partial \mathcal{L}}{\partial w_{kj}} = \sum_m \left[\beta p_k r_{mj} - r_{mj} \exp(w_k^\top r_m) \mathbb{E}_q[\pi_{mk}] \right] - \frac{w_{jk}}{\sigma} \tag{35}$$

$$\frac{\partial^2 \mathcal{L}}{\partial w_{kj'} \partial w_{kj}} = \sum_m \left[-r_{mj'} r_{mj} \exp(w_k^\top r_m) \mathbb{E}_q[\pi_{mk}] \right] - \mathbf{1}(j = j')\sigma^{-1}. \tag{36}$$

Because the w_{kj} is depends on both label j and topic k, we iteratively update w_{kj} until converged.

The update rules for the variational parameter of π_{mk} and V_k need to be changed to accommodate the change of scaling function. The variational parameters of π_{mk} are approximated by using the first-order Taylor expansion,

$$\begin{aligned} a^{\pi}_{mk} &= \beta p_k + \sum_{n=1}^{N_m} \gamma_{mnk} \\ b^{\pi}_{mk} &= \mathbb{E}_q[\ln s(r_m)] + \frac{N_m}{\xi_m}, \end{aligned} \tag{37}$$

where ξ_m is $\sum_{k=1}^{K} \mathbb{E}_q[\pi_{mk}]$. To update V_k, we take the same approach in which the variational distribution is a delta function of current V_k. Again, we use the steepest ascent algorithm to jointly optimize V_k, and the gradient of V_k is

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial V_k} = &- \frac{\alpha - 1}{1 - V_k} - \frac{\beta p_k}{V_k} \left\{ \sum_m \mathbb{E}_q[\ln s(r_m)] \mathbb{E}_q[\ln \pi_{mk}] - \mathbb{E}_q[\pi_{mk}] + \psi(\beta p_k) \right\} \\ &+ \sum_{k' > k} \frac{\beta p_{k'}}{1 - V_k} \left\{ \sum_m \mathbb{E}_q[\ln s(r_m)] \mathbb{E}_q[\ln \pi_{mk'}] - \mathbb{E}_q[\pi_{mk'}] + \psi(\beta p_{k'}) \right\}. \end{aligned} \tag{38}$$

The update rules for π_{mk} and V_k only requires the expectation of the log scaling function. The update rules for the other parameters remain the same as the previous section.

There might be possible alternatives for a scaling function with respect to characteristics of dataset used. Introducing a new scaling function requires a new inference algorithm, and this can be cumbersome. Recently, several approaches have been proposed to bypass the complex derivation of variational updates (Kingma and Welling 2014; Ranganath et al. 2014; Tran et al. 2016). Most of these approaches rely on re-parameterization tricks and stochastic updates with random samples from variational distributions. Although these methods are unbiased estimators of the variational parameters, sometimes they suffer from high variance of the samples, especially, when they are applied for the whole ELBO (Ranganath et al. 2014). We suggest to infer the scaling irrelevant parameters using the provided variational updates and scaling relevant parameters using these black-box techniques to reduce the possible high variances of these approaches.

5 Experiments

In this section, we describe how the HDSP performs with real and synthetic data. We fit the HDSP topic model with three different types of data and compare the results with several comparison models. First, we test the model with synthetic data to verify the approximate inference. Second, we train the model with categorical data whose label information is represented by binary values. Third, we train the model with mixed-type of data whose label information has both numerical and categorical values.

5.1 Synthetic data

There is no naturally-occurring dataset with the observable weights between topics and labels, so we synthesize data based on the model assumptions to verify our model and the approximate inference. First, we check the difference between the original topics and the inferred topics via simple visualization. Then, we focus on the differences between the inferred and synthetic weights. For all experiments with synthetic data, the datasets are generated by following the model assumptions with the first scaling function, and the posterior inferences are done with the first scaling function. We set the truncation level T at twice the number of topics. We terminate the variational inference when the fractional change of the lower bound falls below 10^{-3}, and we average all results over 10 individual runs with different initializations.

With the first experiment, we show that HDSP correctly recovers the underlying topics and scaling parameter between topics and labels. For the dataset, we generate 500 documents using the following steps. We define five topics over ten terms shown in Fig. 2a and the scaling parameter of five topics and four labels shown in Fig. 2d. For each document, we randomly draw N_m from the Poisson distribution and r_{mj} from the Bernoulli distribution. The average length of a document is 20, and the average number of labels per document is 2. We generate topic proportions of corpus and documents by using Eqs. 10 and 13. For each word in a document, we draw the topic and the word by using Eq. 14. We set both α and β to 1.

Figure 2 shows the results of the HDP and the HDSP on the synthetic dataset. Figure 2b, c are the heat maps of topics inferred from each model. We match the inferred topics to the original topics using KL divergence between the two sets of topic distributions. There are no significant differences between the inferred topics of HDSP and HDP. In addition to the topics, HDSP infers the scaling parameters between topics and labels, which are shown in Fig. 2e. The results show that the relative differences between original scaling parameters are preserved in the inferred parameters through the variational inference.

With the second experiment, we show that the inferred parameters preserve the relative differences between labels and topics in the dataset. For this experiment, we generate 1,000 documents with ten randomly drawn topics from Dirichlet(0.1) with the vocabulary size of 20. To generate the weights between topics and labels, we randomly place the topics and labels into three dimensional euclidean space, and use the distance between a topic and label as a scaling parameter. Let $\theta_k \in \mathbb{R}^3$ be a location of topic k and $\theta_j \in \mathbb{R}^3$ be a location of label j. We use $|\theta_k - \theta_j|_2$ as an inverse scaling parameter w_{kj}^{-1} between topic k and label j, so the scaling weight increase as a distance between a topic and a label decreases. The location of topics and labels are uniformly drawn from three dimensional euclidean space, so the total volume is x^3, then we vary the x value from 1 to 20 for each experiment.

We compute the mean absolute error (MAE) and the spearman's rank correlation coefficient ρ between the original parameters and the inferred parameters. The spearman's ρ is designed to measure the ranking correlation of two lists. Figure 3 shows the results. The MAE increases as the volume of the space increases. However, spearman's ρ stabilizes, indicating that the relative differences are preserved even when the MAE increases. Since there are an infinite number of configurations of scaling parameters that generate the same expectation

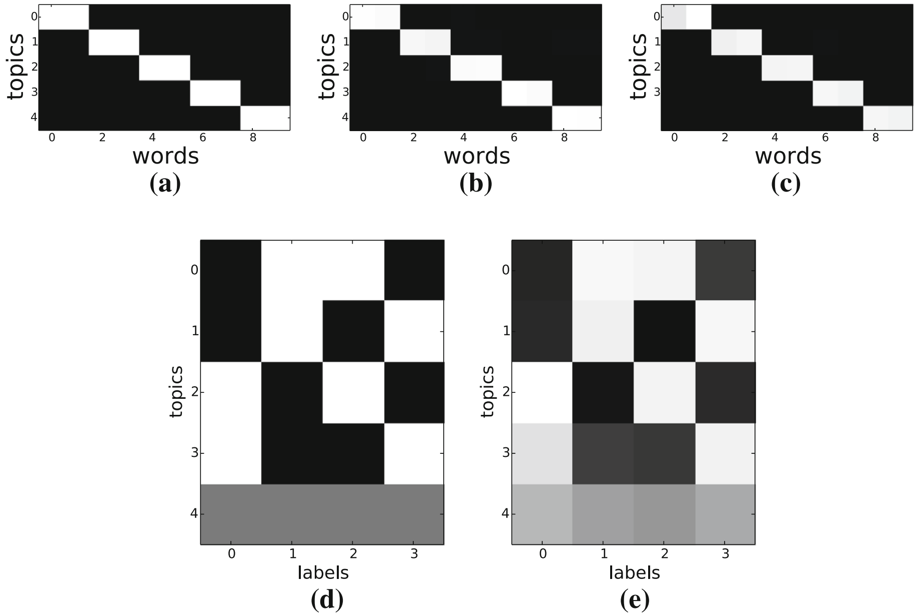


Fig. 2 Experiments with synthetic data. **a** Is the synthetic topic distribution of 5 topics over 10 terms. **b**, **c** Are topic distributions inferred by the HDSP and the HDP. Both models recover the original topics. **d** Shows the heat map of original scaling parameters between the topics and labels. **e** Shows the heat map of the recovered parameters by HDSP

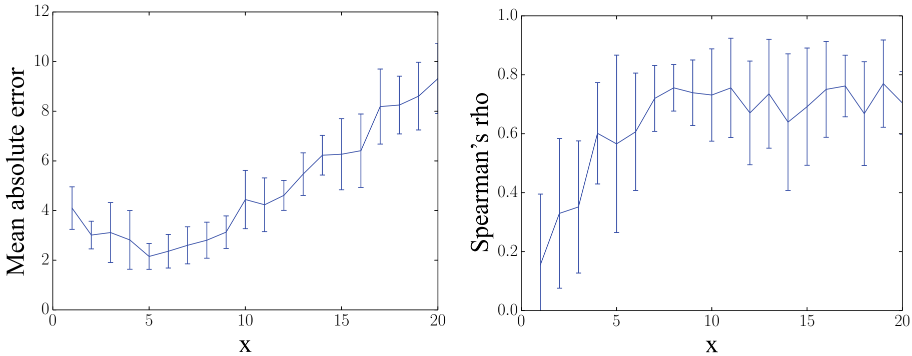
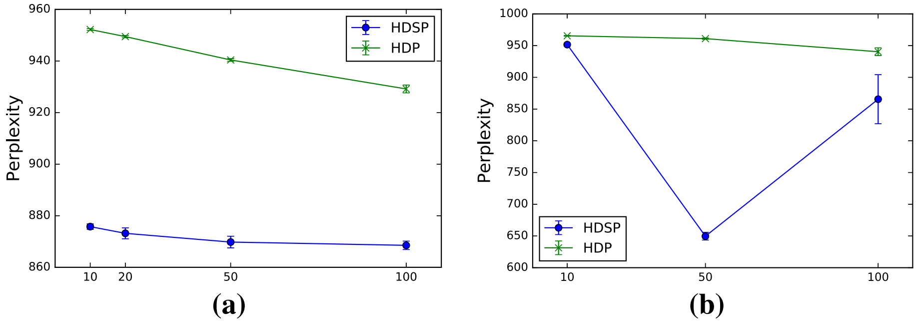
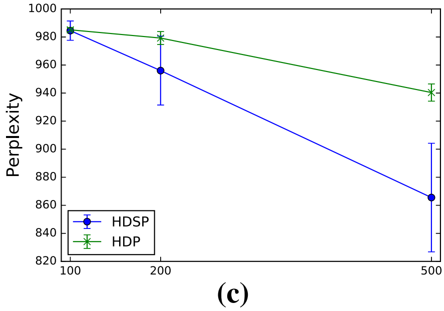


Fig. 3 Spearman's correlation coefficient and mean absolute error of the synthetic data with various volume of space (x^3). As the volume of space for locations increases, the mean absolute error also increases (*left*). However, the model preserves the relative weights between topics and labels, shown by the high and stabilized correlation between the original ordering and the recovered ordering of label-topic pairs in terms of the weights between the two (*right*). This is a key characteristic of the HDSP model which scales the mixture components according to the inverse of the weights

$\mathbb{E}[p(\pi_m|\beta p, w_j)]$ given π_m and βp, preserving the relative differences verifies our model's capability of capturing the underlying structure of topics and labels.

Finally, we compare the predictive performance of HDSP to that of HDP to understand where the covariate information can help guess the topics. For this experiment, we generate additional synthetic documents in the same way used for the previous experiment. The detailed parameter settings are shown in the table in Fig. 4. In addition, we vary the length





Parameter	value
# of documents	500
# of labels	100
vocabulary size	1,000
# of topics	200
document length	50

Fig. 4 Perplexity on heldout synthetic documents with various settings of document length, number of labels, and number of documents. Table shows the default parameters used to generate the synthetic datasets. For each experiment, we vary the value of a single parameter while fixing the rest. **a** Length of a document, **b** number of labels, **c** number of documents

of documents, the number of labels, and the number of documents while the rest of the parameters are fixed. To compare the predictive performance of the two models, we compute expected perplexity (Eq. 41) on the test set given label information. Lower perplexity indicates better performance. Figure 4 reveals some characteristics of HDSP. First, HDSP shows relatively lower perplexity when the length of documents is short. Although the perplexity of HDSP is consistently lower than that of HDP in Fig. 4a, the differences are larger when the length of each document is relatively short. Second, the number of labels affects the predictive performance. Figure 4b shows the different numbers of labels, 10, 50 and 100 labels. HDSP and HDP perform similarly for the 10-label setting, while HDP performs significantly better for 50 and 100 labels. The difference is smaller for the 100-label setting, and we conjecture that this is because the number of labels is too large for the number of training documents. Finally, Fig. 4c shows that with respect to the number of training documents, the perplexity of HDSP decreases as the number of training documents increases.

5.2 Categorical data

We evaluate the performance of HDSP and compare it with the HDP, labeled LDA (L-LDA), partially labeled LDA (PLDA), and author-topic model (ATM). For the HDSP, we use both scaling functions and denote the model with the second scaling function as wHDSP. We use three multi-labeled corpora: RCV[2] (newswire from Reuter's), OHSUMED[3] (a subset of

[2] http://trec.nist.gov/data/reuters/reuters.html.

[3] http://ir.ohsu.edu/ohsumed/ohsumed.html.

Table 1 Datasets used for the experiments in 5.2

	Docs	Vocab	Labels	Labels/doc	Doc/labels
RCV	23,149	9911	117	3.2	729.7
OHSUMED	7505	7056	52	5.2	722.0
NIPS	2484	14,036	2865	2.4	1.6

As the last two columns show, we experiment on datasets with a varied number of unique labels, as well as the average number of labels per document

the Medline journal articles), and NIPS (proceedings of NIPS conference). For RCV and OHSUMED, we use multi-category information of documents as labels, and for NIPS, we use authors of papers as labels. The average number of labels per article is 3.2 for RCV, 5.2 for OHSUMED, and 2.4 for NIPS. Table 1 contains the details of the datasets.

5.2.1 Experimental settings

For the HDP and HDSP, we initialize the word-topic distribution with three iterations of LDA for fast convergence to the posterior while preventing the posterior from falling into a local mode of LDA and then reorder these topics by the size of the posterior word count. For all experiments, we set the truncation level T to 200. We terminate variational inference when the fractional change of the lower bound falls below 10^{-3}, and we optimize all hyperparameters during inference except η. For the L-LDA and PLDA, we implement the collapsed Gibbs sampling algorithm. For each model, we run 5000 iterations, the first 3000 as burn-in and then using the samples thereafter with gaps of 100 iterations. For PLDA, we set the number of topics for each label to two and five (PLDA-2, PLDA-5). For the ATM, we set the number of topics to 50, 100, and 150. We try five different values for the topic Dirichlet parameter η: $\eta = 0.1, 0.25, 0.5, 0.75, 1.0$. Finally all results are averaged over 20 runs with different random initializations. We do not report the standard errors because they are small enough to ignore.

5.2.2 Evaluation metric

The goal of our model is to construct the dependent random probability measure given multiple labels. Therefore, our interest is to see the increments of predictive performance when the label information is given.

The predictive probability given label information for held-out documents are approximated by the conditional marginal,

$$p(\mathbf{x}'|\mathbf{r}', \mathcal{D}_{\text{train}}) = \int_q \prod_{n=1}^{N} \sum_{k=1}^{T} p(\mathrm{x}'_n|\phi_k) p(z'_n = k|\pi') p(\pi'|V, \mathbf{r}') dq(V, w, \phi), \tag{39}$$

where $\mathcal{D}_{\text{train}} = \{\mathbf{x}_{\text{train}}, \mathbf{r}_{\text{train}}\}$ is the training data, $\mathbf{x}'$ is the vector of N words of a held-out document, $\mathbf{r}'$ are the labels of the held-out document, z'_n is the latent topic of word n, and π'_k is the kth topic proportion of the held-out document. Since the integral is intractable, we approximate the probability

$$p(\mathbf{x}'|\mathbf{r}', \mathcal{D}_{\text{train}}) \approx \prod_{n=1}^{N} \sum_{k=1}^{T} \tilde{\pi}_k \tilde{\phi}_{k,\mathrm{x}'_n}, \tag{40}$$

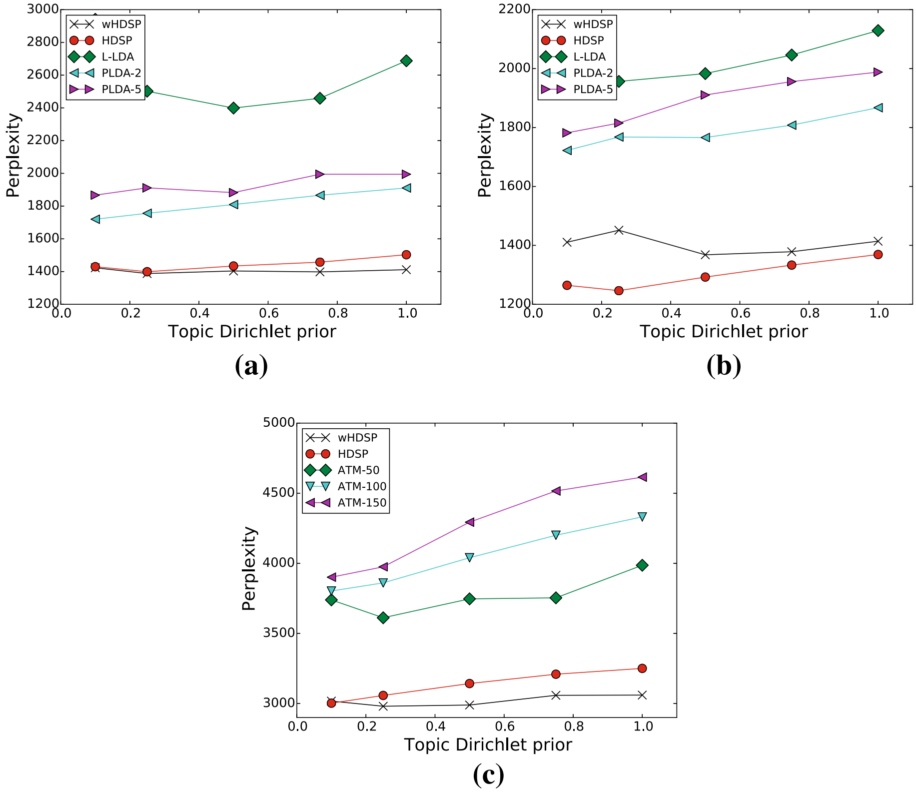


Fig. 5 Perplexity of held-out documents. For HDSP (first scaling function), wHDSP (second scaling function), L-LDA, ATM, and PLDA, the perplexity is measured given documents and observed labels. For HDP, the model only uses the words of the documents. The HDSP which, instead of excluding topics from unobserved labels, scales all topics according to observed labels, shows the best heldout perplexity. **a** OHSUMED, **b** RCV, **c** NIPS

where $\tilde{\phi}_k$ and $\tilde{\pi}_k$ are the variational expectations of ϕ_k and π_k given label $\mathbf{r}'$. This approximated likelihood is then used to compute the perplexity of the held-out document

$$\text{perplexity} = \exp\left\{\frac{-\ln p(\mathbf{x}'|\mathbf{r}', \mathcal{D}_{\text{train}})}{N}\right\}. \tag{41}$$

Lower perplexity indicates better performance. We also take the same approach to compute the perplexity for L-LDA, and PLDA. To measure the predictive performance, we leave 20% of the documents for testing and use the remaining 80% to train the models.

5.2.3 Experimental results

Figure 5 shows the predictive performance of our model against the comparison models. For the OHSUMED and RCV corpora, both HDSP and wHDSP outperform all others. Among these models, L-LDA restricts the modeling flexibility the most; the PLDA relaxes that restriction by adding an additional latent label and allowing multiple topics per label. HDSP and wHDSP further increase the modeling flexibility by allowing all topics to be generated from each label. This is reflected in the results of predictive performance of the three models;

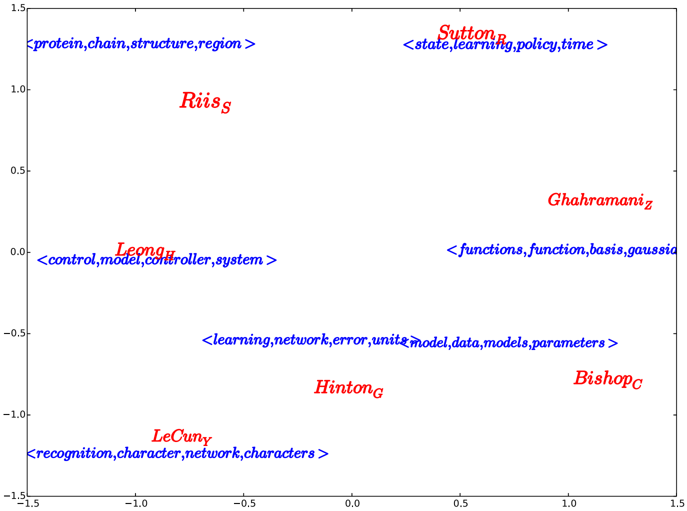


Fig. 6 Relative locations of observed labels (*red*) and latent topics (*blue*) inferred by HDSP from the NIPS corpus

L-LDA shows the worst performance, then PLDA, and HDSP and wHDSP show the lowest perplexity. For the NIPS data, we compare HDSP and wHDSP to ATM, and again, HDSP and wHDSP show the lowest perplexity.

To visualize the relationship between topics and labels, we embed the inferred topics and the labels into the two dimensional euclidean space by using multidimensional scaling (Kruskal 1964) on the inferred parameters of HDSP. In Fig. 6, we choose and display a few representative topics and authors from NIPS. For instance, Geoffrey Hinton and Yann LeCun are closely located to the neural network related topics such as ‘learning, network error’ and ‘recognition character, network’, and the reinforcement learning researcher Richard Sutton is closely located to the ‘state, learning policy’ topic. Figure 7 shows the embedded labels and topics from OHSUMED. The labels ‘Preschool’, ‘Pregnancy’, and ‘Infant’ are closely located to one another with similar topics. While the model explicitly models the correlation between topics and labels, embedding them together shows that the correlation among labels, as well as among topics, can also be inferred. Additional visualizations on relationship between topics and labels are provided in “Appendix 3”.

5.2.4 Modeling data with missing labels

We also test our model with partially labeled data which have not been previously covered in topic modeling. Many real-world data fall into this category where some of the data are labeled, others are incompletely labeled, and the rest are unlabeled. For this experiment, we randomly remove existing labels from the RCV and OHSUMED corpora. To remove observed labels in the training corpus, we use Bernoulli trials with varying parameters to analyze how the proportion of observed labels affects the heldout predictive performance of the model.

Figure 8 shows the predictive perplexity with varying parameters of Bernoulli distribution from 0.1 to 0.9. For both scaling functions, the perplexity decreases as the model observes more labels. Compared to the PLDA (with the parameter setting for optimal performance),

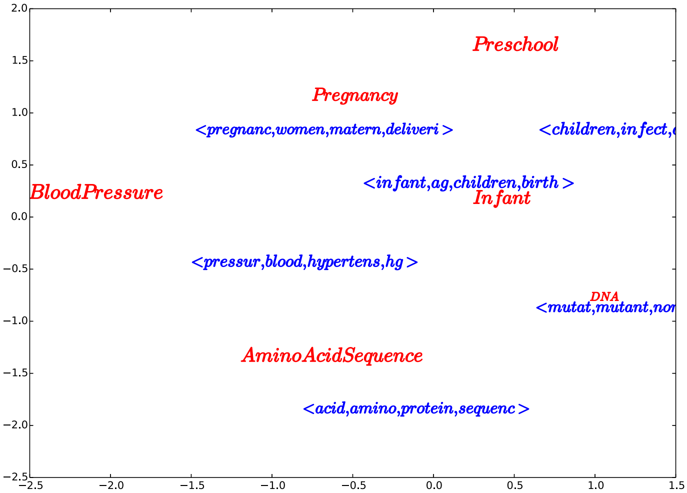


Fig. 7 Relative locations of observed labels (*red*) and latent topics (*blue*) inferred by HDSP from the OHSUMED corpus

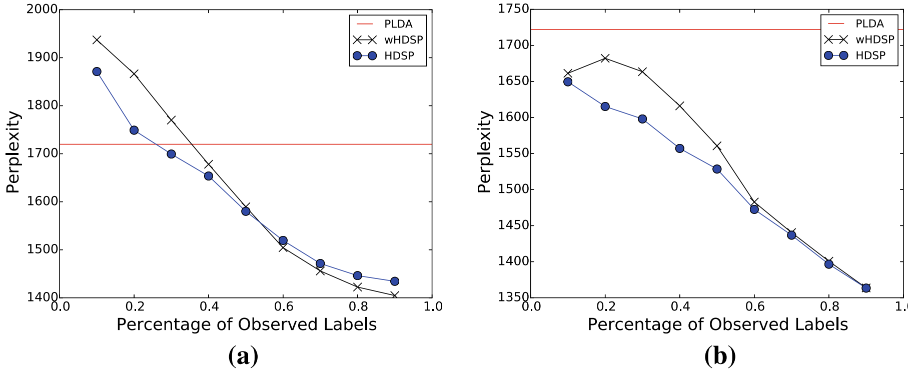


Fig. 8 Perplexity of partially labeled documents. For both RCV and OHSUMED, we randomly remove the observed labels of training documents based on Bernoulli trials. After training the model with removed dataset, we measure the heldout perplexity on a test documents with different scaling functions. **a** OHSUMED, **b** RCV

the HDSP achieves similar perplexity with only 20% of the labels. One notable phenomenon is that the HDSP outperforms wHDSP on both datasets when the number of observed labels is less than 50% of the total number of labels.

5.2.5 Modeling data with single category

The HDSP has been designed to model multi-labeled documents but it can also be used for single-labeled documents. In this section, we measure the classification performance of the HDSP on a single label corpus. To measure the classification performance, we trained our model with five comp subcategories of newsgroup documents (20NG). Table in Fig. 9 shows the details of this dataset. 90% of the documents were used with the labels, and the remaining 10% of documents were used without the labels. We classified each of the test documents by the label with the lowest expected perplexity given that label. As a baseline,

20NG dataset	
# of documents	1,800
# of labels	5
vocabulary size	2,608

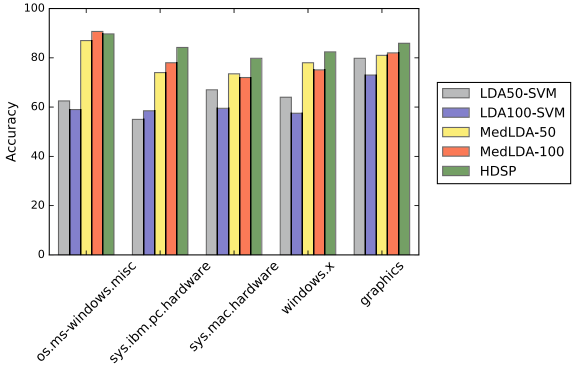


Fig. 9 Data statistics of 20NG dataset, and accuracies of HDSP, MedLDA, and LDA-SVM on classification of 20NG dataset. HDSP outperforms LDA-SVM for all five labels

we trained a multi-class SVM with the topic proportions inferred by LDA (LDA-SVM). We also compared the results of HDSP with MedLDA (Zhu et al. 2009), a supervised topic model. The results, shown in Fig. 9, display a significant improvement of our model over the LDA-SVM approach and MedLDA.

5.3 Mixed-type data

In this section, we present the performance of the second scaling function with a corpus of product reviews which has real-valued ratings and category information.

The first scaling function is only applicable to categorical side information, so we use the second scaling function (wHDSP) which can model numerical as well as categorical side information of documents. To evaluate the performance of wHDSP with numerical side information, we train the model with the Amazon review data collected from seven categories of electronic products: air conditioner, canister vacuum, coffee machine, digital SLR, laptop, MP3 player, and space heater. Amazon uses a five-star rating system, so each review contains one numerical rating ranging from one to five. Table 2 shows the number of reviews for each rating and category. Recall that r is a vector whose values denote the observation of the labels. For each review, we set the dimension of r to eight in which the first dimension is a numerical rating of a review, and then the remaining seven dimensions match the seven product categories. We set the value of each dimension to one if the review belongs to the corresponding category, and zero otherwise.

To evaluate the performance of wHDSP, we classify the ratings of the reviews based on a trained model. We use 90% of the corpus to train models and the remaining 10% of the corpus to test the models. To classify the rating of each review in the test set, we compute the perplexity of the given review with varying ratings from one to five, and choose the rating that shows the lowest perplexity. Generally, computing the perplexity of heldout document requires complex approximation schemes (Wallach et al. 2009), but we compute the perplexity based on the expected topic distribution given category and rating information, which requires a finite number of computations.

We compare the wHDSP with the supervised LDA (SLDA), LDA + SVM, as well as classifiers Naive Bayes, SVM, and decision trees (CART). For the LDA + SVM approach, we first train the LDA model and then use the inferred topic proportion and categories as features of the SVM. For the SLDA model, the category information cannot be used because

Table 2 The number of reviews for each rating and category in the Amazon dataset

	# reviews	Percentage
Total	24,259	100
5-star	12,382	52
4-star	5040	20
3-star	1905	8
2-star	1723	7
1-star	3209	13
Category		**# reviews**
Canister vacuum		3535
Digital SLR		4189
Laptop		4252
MP3		3659
Air conditioner		568
Space heater		3859
Coffee machine		4197

The dataset contains 24,259 reviews collected from seven different product categories. The description reveals a highly skewed distribution of reviews where 52% of reviews are rated as five-star. The percentage denotes the proportion of reviews per each rating

the model is designed to learn and predict the single response variable. For both models, we set the number of topics to 50, 100, and 200.

In many applications, classifying negative feedback of users is more important than classifying positive feedback. From the negative feedback, companies can identify possible problems of their products and services and use the information to design their next product or improve their services. In most online reviews, however, the proportion of negative feedback is smaller than the proportion of positive feedback. For example, in the Amazon data, about 51% of reviews are rated as five-star, and 72% rated as four or five. A classifier trained by such skewed data is likely to be biased toward the majority class.

We report the classification results of Amazon dataset in Tables 3 and 4. Table 3 shows the results for each rating in terms of F1, and Table 4 shows the results in terms of micro and macro F1. The wHDSP outperforms the other models in terms of macro F1 but performs worse than sLDA in terms of micro F1. As we noted earlier, classifying negative reviews may be more important in many applications. Both the SLDA with 100 topics and the wHDSP are comparable in classifying the most negative (one-star) reviews. However, the confusion matrices and Table 3 indicate that the SLDA dichotomously learns the decision boundaries where the most reviews are classified into one-star or five-star. For example, the SLDA with 50 and 100 topics did not classify any two-star and three-star reviews correctly. The wHDSP learns the decision boundaries for classifying subtle differences between five-star rating reviews. These patterns are shown clearly with the confusion matrices in Fig. 10 where the diagonal entries are the numbers of correctly classified reviews. For example, the wHDSP classified only eight one-star reviews as five-star reviews, but the SLDA50 assigned 68 reviews as five-star reviews.

We perform a rating prediction task with and without the category information of reviews to see the effect of using both the category and rating information on the wHDSP and LDA + SVM approaches. The results represented by wHDSP* in Table 4 and Fig. 10b show the performance of rating prediction with the wHDSP trained without category information. For wHDSP*, the model performs worse than wHDSP, which indicates the model, without category information, cannot distinguish the review ratings which depend on topical context.

Table 3 F1 of wHDSP and the other models for the Amazon review corpus

F1	Ratings				
	1	2	3	4	5
wHDSP	0.600	**0.161**	**0.185**	**0.316**	0.687
wHDSP-no-cate	0.428	0.087	0.099	0.061	0.658
LDA50 + SVM	0.392	0.036	0.038	0.134	0.684
LDA100 + SVM	0.454	0.078	0.073	0.265	0.678
LDA200 + SVM	0.508	0.032	0.100	0.284	0.681
SLDA50	0.603	0.000	0.021	0.140	**0.741**
SLDA100	0.606	0.000	0.021	0.067	0.740
SLDA200	0.580	0.015	0.011	0.140	0.727
SVM	0.403	0.000	0.000	0.007	0.716
NaiveBayes	**0.634**	0.028	0.085	0.469	0.652
DecisionTree	0.457	0.088	0.154	0.355	0.628

wHDSP and SLDA perform comparably on one-star ratings but wHDSP outperforms SLDA on middle range ratings (two, three, and four stars)

Table 4 Macro and micro F1 of the wHDSP and the other models

5-Ratings	MacroF1	MicroF1
wHDSP	**0.390**	0.522
wHDSP*	0.267	0.474
LDA50 + SVM	0.257	0.518
LDA100 + SVM	0.310	0.520
LDA200 + SVM	0.321	0.527
LDA200 + SVM*	0.309	0.533
SLDA50	0.301	0.584
SLDA100	0.287	**0.588**
SLDA200	0.294	0.577
SVM	0.225	0.560
NaiveBayes	0.374	0.545
DecisionTree	0.336	0.477

The first column shows the performance of classification with five-star ratings system. The models with asterisk are trained without category information. Note that the SLDA cannot incorporate two different types of labels together

The LDA + SVM without categories achieves 0.309 macro F1 and 0.533 micro F1, which are comparable to the LDA + SVM with the category information. Unlike the wHDSP, the decision boundaries of SVM are not improved with the additional category information. The result supports that for learning decision boundaries between ratings over different categories, the approach of including category information to train topics is more effective than using topics and the category information independently.

6 Discussions

We have presented the hierarchical Dirichlet scaling process (HDSP), a Bayesian nonparametric prior for a mixed membership model that lets us analyze underlying semantics and observable side information. The combination of the stick breaking process with the normalized gamma process in HDSP is a more controllable construction of the hierarchical Dirichlet process because each atom of the second level measure inherits from the first level measure in order. HDSP also allows more flexibility and the capability of modeling side information

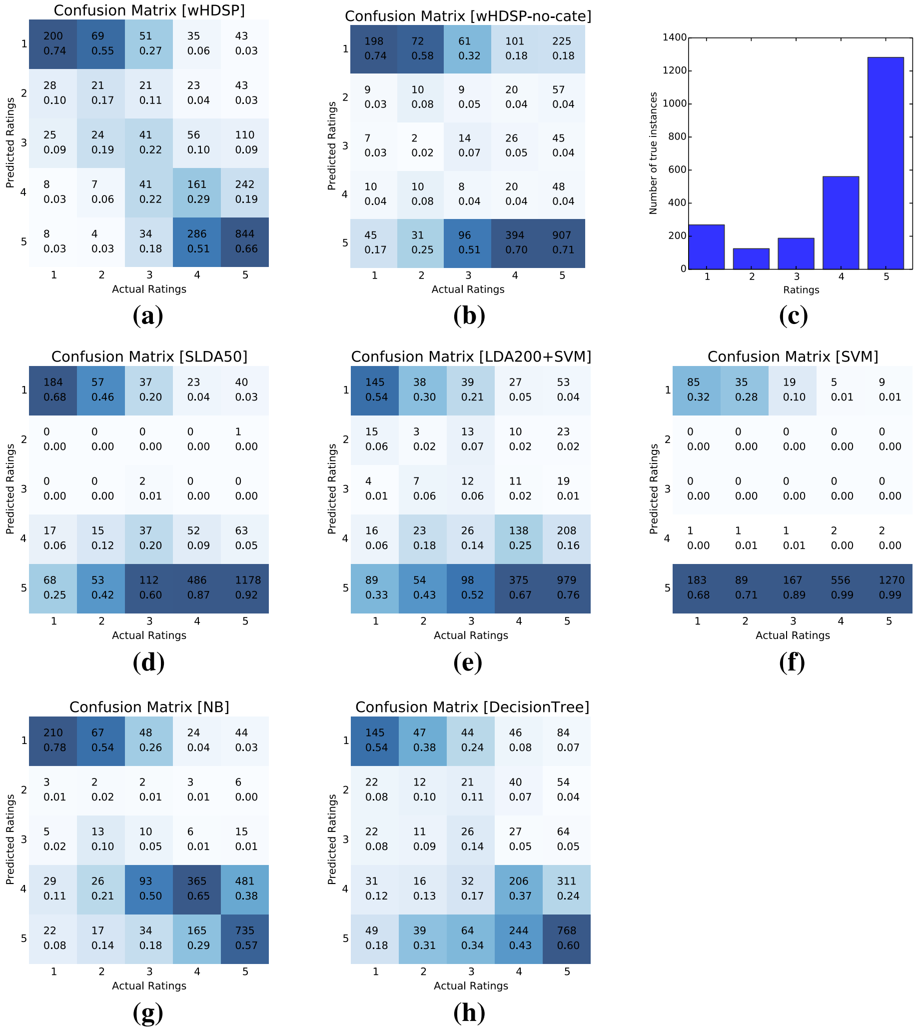


Fig. 10 Confusion matrices from classification results of wHDSP and the other models. *Diagonal entries* indicate the number of correctly classified reviews per rating. Despite the highly skewed data distribution (**c**), wHDSP achieves relatively better classification results for negative and neutral reviews as shown in (**a**). Confusion matrices are column-wise normalized. **a** wHDSP, **b** wHDSP without categories, **c** # of instances per rating, **d** SLDA50, **e** LDA200 + SVM, **f** SVM, **g** NAvieBayes, **h** DecisionTree

by the scaling functions that plug into the rate parameter of the gamma distribution. The choice of the scaling function is the most important part of the model in terms of establishing a link between topics and observed labels. We developed two scaling functions but the choice of scaling function depends on the modeler's intention. For example, the well known linking functions from the generalized linear model can be used as scaling functions, or one can use several scaling functions together on purpose. We showed that the application of HDSP to topic modeling correctly recovers the topics and topic-label weights of synthetic data. Experiments with the real dataset show that the first scaling function is more suited for partially labeled data, and the second scaling function is more suited for a dataset with both numerical and categorical labels.

Hierarchical Dirichlet scaling process opens up a number of interesting research questions that should be addressed in future work. First, in the two scaling functions we proposed to model the correlation structure between topics and side information, we simply defined the relationship between topic k and label j through the scaling parameter w_{kj}. However, this approach does not consider the correlation within topics and labels. Taking inspiration from previous work (Blei and Lafferty 2007; Mimno et al. 2007; Paisley et al. 2012) that showed correlations among topics, we can define a scaling function with a prior over the topics and labels to capture their complex relationships. Second, our posterior inference algorithm based on mean-field variational inference is tested with tens of thousands documents. However, modern data analysis requires inference of massive and/or streaming data. For a fast and efficient posterior inference, we can apply parallel or distributed algorithms based on a stochastic update (Hoffman et al. 2013; Ahn et al. 2014). Furthermore, we fix the number of labels before training but we need to find a way to model the unbounded number of labels for streaming data.

Acknowledgements This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government (MSIP) (No.B0101-15-0307, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Center)).

Appendix 1: Variational inference for HDSP

In this section, we provide the detailed derivation for mean-field variational inference for HDSP with the first scaling function. First, the evidence of lower bound for HDSP is obtained by taking a Jensen's inequality on the marginal log likelihood of the observed data,

$$\ln \int p(\mathcal{D}, \Theta) d\Theta \geq \int Q(\Psi) \ln \frac{P(\mathcal{D}, \Theta)}{Q(\Psi)} d\Theta, \tag{42}$$

where $\mathcal{D}$ is the training set of documents and labels, Ψ denotes the set of variational parameters, and Θ denotes the set of model parameters.

We define a fully factorized variational distribution Q as follows:

$$Q := \prod_{k=1}^{T} q(\phi_k) q(V_k) \prod_{j=1}^{J} q(w_{kj}) \prod_{m=1}^{M} q(\pi_{mk}) \prod_{n=1}^{N_m} q(z_{mn}), \tag{43}$$

where

$$\begin{aligned} q(z_{mn}) &= \text{Multinomial}(z_{mn} | \gamma_{mn1}, \gamma_{mn2}, \ldots, \gamma_{mnT}) \\ q(\pi_{mk}) &= \text{Gamma}(\pi_{mk} | a^{\pi}_{mk}, b^{\pi}_{mk}) \\ q(w_{kj}) &= \text{InvGamma}(w_{kj} | a^{w}_{kj}, b^{w}_{kj}) \\ q(\phi_k) &= \text{Dirichlet}(\phi_k | \eta_{k1}, \eta_{k2}, \ldots, \eta_{kI}) \\ q(V_k) &= \delta_{V_k}. \end{aligned} \tag{44}$$

The evidence of lower bound (ELBO) is

$$L(\mathcal{D}, \Psi) = \mathbb{E}_q[\ln p(\mathcal{D}, \Psi)] + \mathbb{H}[Q]$$

$$
\begin{aligned}
&= \mathbb{E}_q\left[\sum_{m=1}^{M}\sum_{n=1}^{N_m} \ln p(\mathrm{x}_{mn}|z_{mn}, \Phi)\right] + \mathbb{E}_q\left[\sum_{m=1}^{M}\sum_{n=1}^{N_m} \ln p(z_{mn}|\pi_m)\right] \\
&\quad + \mathbb{E}_q\left[\sum_{m=1}^{M}\sum_{k=1}^{\infty} \ln p(\pi_{mk}|V_k, w_k, \mathbf{r}_m)\right] + \mathbb{E}_q\left[\sum_{k=1}^{\infty} \ln p(V_k|\alpha)\right] \\
&\quad + \mathbb{E}_q\left[\sum_{j=1}^{J}\sum_{k=1}^{\infty} \ln p(w_{kj}|a^w, b^w)\right] \\
&\quad + \mathbb{E}_q\left[\sum_{k=1}^{\infty} \ln p(\phi_k|\eta)\right] - \mathbb{E}_q[\ln Q] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{T} \gamma_{mnk}\mathbb{E}_q[\ln p(\mathrm{x}_{mn}|\phi_k)] + \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{T} \gamma_{mnk}\mathbb{E}_q[\ln p(z_{mn} = k|\pi_m)] \\
&\quad + \sum_{m=1}^{M}\sum_{k=1}^{T} \mathbb{E}_q[\ln p(\pi_{mk}|V_k, w_k, \mathbf{r}_m)] + \sum_{k=1}^{T} \mathbb{E}_q[\ln p(V_k|\alpha)] \\
&\quad + \sum_{k=1}^{\infty}\sum_{j=1}^{J} \mathbb{E}_q[\ln p(w_{kj}|a^w, b^w)] - \mathbb{E}_q[\ln Q]
\end{aligned} \tag{45}
$$

where the expectations of latent variables under the variational distribution Q are

$$
\begin{aligned}
&\mathbb{E}_q[\pi_{mk}] = a^{\pi}_{mk}/b^{\pi}_{mk} \\
&\mathbb{E}_q[\ln \pi_{mk}] = \psi(a^{\pi}_{mk}) - \ln b^{\pi}_{mk} \\
&\mathbb{E}_q[w_{kj}] = b^{w}_{kj}/(a^{w}_{kj} - 1) \\
&\mathbb{E}_q[w^{-1}_{kj}] = a^{w}_{kj}/b^{w}_{kj} \\
&\mathbb{E}_q[\ln w_{kj}] = \ln b^{w}_{kj} - \psi(a^{w}_{kj}) \\
&\mathbb{E}_q[\ln \phi_{ki}] = \psi(\eta_{kd}) - \psi(\sum_{i}^{'} \psi_{ki'})
\end{aligned}
$$

Then, we derive the equations further

$$
\begin{aligned}
L(\mathcal{D}, \Psi) &= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{T} \gamma_{mnk}\left\{\psi(\eta_{k\mathrm{x}_{mn}}) - \psi(\sum_{d} \eta_{kd})\right\} \\
&\quad + \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{T} \gamma_{mnk}\left\{\mathbb{E}_q[\ln \pi_{mk}] - \mathbb{E}_q\left[\ln \sum_{k=1}^{T} \pi_{mk}\right]\right\} \\
&\quad + \sum_{m=1}^{M}\sum_{k=1}^{T} -\beta p_k \sum_{j} r_{mj}\{\ln(b^{w}_{kj}) - \psi(a^{w}_{kj})\} + (\beta p_k - 1)\{\psi(a^{\pi}_{mk}) - \ln(b^{\pi}_{mk})\} \\
&\quad - \prod_{j}\left(\frac{a^{w}_{kj}}{b^{w}_{kj}}\right)^{r_{mj}} \frac{a^{\pi}_{mk}}{b^{\pi}_{mk}} - \ln \Gamma(\beta p_k)
\end{aligned}
$$

$$+\sum_{k=1}^{T}\ln\Gamma(\alpha+1)-\ln\Gamma(\alpha)+(\alpha-1)\ln(1-V_k)$$
$$+\sum_{k=1}^{T}\sum_{j=1}^{J}a^{w}\ln b^{w}-\ln\Gamma(a^{w})-(a^{w}+1)\{\ln b^{w}-\psi(a^{w})\}-a^{w}-\mathbb{E}_q[\ln Q]. \tag{46}$$

Taking the derivatives of this lower bound with respect to each variational parameter, we can obtain the coordinate ascent updates.

The optimal form of the variational distribution can be obtained by exponentiating the variational lower bound with all expectations except the parameter of interest (Bishop and Nasrabadi 2006). For π_{mk}, we can derive the optimal form of variational distribution as follows

$$\begin{aligned} q(\pi_{mk}) &\propto \exp\left\{\mathbb{E}_{q_{-\pi_{mk}}}\left[\ln p\,(\pi_{mk}|z,\pi_{-mk},\mathbf{r}_m,V)\right]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q_{-\pi_{mk}}}\left[\ln p(z|\pi_m)+\ln p(\pi_m|\mathbf{r}_m,V)\right]\right\} \\ &\propto \pi_{mk}^{\beta p_k+\sum_{n=1}^{N_m}\gamma_{mnk}-1}e^{(b^{\pi}_{mk}\prod_j \mathbb{E}_q[w_{kj}^{-r_{mj}}]+\frac{N_m}{\xi_m})\pi_{mk}} \end{aligned} \tag{47}$$

where update for ξ_m is $-\ln\xi_m-(\sum_{k=1}^{T}\mathbb{E}_q[\pi_{mk}]-\xi_m)/\xi_m$. Therefore, the optimal form of variational distribution for π_{mk} is

$$q(\pi_{mk})\sim\text{Gamma}\left(\beta p_k+\sum_{n=1}^{N_m}\gamma_{mnk},\quad b^{\pi}_{mk}\prod_j\mathbb{E}_q[w_{kj}^{-r_{mj}}]+\frac{N_m}{\xi_m}\right). \tag{48}$$

We take the same approach described in (Paisley et al. 2012), and the only difference comes from the product of the inverse distance term.

For w_{kj}, we can derive the optimal form of the variational distribution as follows

$$\begin{aligned} q(w_{kj}) &\propto \exp\left\{\mathbb{E}_{q_{-w_{kj}}}[\ln p(w_{kj}|\pi,w_{-jk},a^{w},b^{w})]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q_{-w_{kj}}}\left[\sum_{m=1}^{M}\ln p\left(\pi_{mk}|\beta p_k,\prod_{j=1}^{J}w_{kj}^{-r_{mj}}\right)+\ln p\left(w_{kj}|a^{w},b^{w}\right)\right]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q_{-w_{kj}}}\left[-\beta p_k\sum_{m=1}^{M}r_{mj}\ln w_{kj}-\sum_m\prod_j w_{kj}^{-r_{mj}}\pi_{mk}-(a^{w}+1)\ln w_{kj}-\frac{b^{w}}{w_{kj}}\right]\right\} \\ &\propto w_{kj}^{-\mathbb{E}_q[\beta p_k]\sum_{m=1}^{M}r_{mj}-a^{w}-1}e^{(-\sum_{\{m:r_{mj}=1\}}\prod_{\{j':r_{mj'}=1/j\}}\mathbb{E}_q[w_{j'k}^{-1}]\mathbb{E}_q[\pi_{mk}]-b^{w})\frac{1}{w_{kj}}} \end{aligned} \tag{49}$$

Therefore, the optimal form of variational distribution for w_{kj} is

$$q(w_{kj})\sim\text{InvGamma}(\mathbb{E}_q[\beta p_k]\sum_m r_{mj}+a^{w},\quad \sum_{m'}\prod_{j'/j}\mathbb{E}_q[w_{j'k}^{-1}]\mathbb{E}_q[\pi_{m'k}]+b^{w}) \tag{50}$$

where $m'=\{m:r_{mj}=1\}$ and $j'/j=\{j':r_{mj'}=1,j'\neq j\}$.

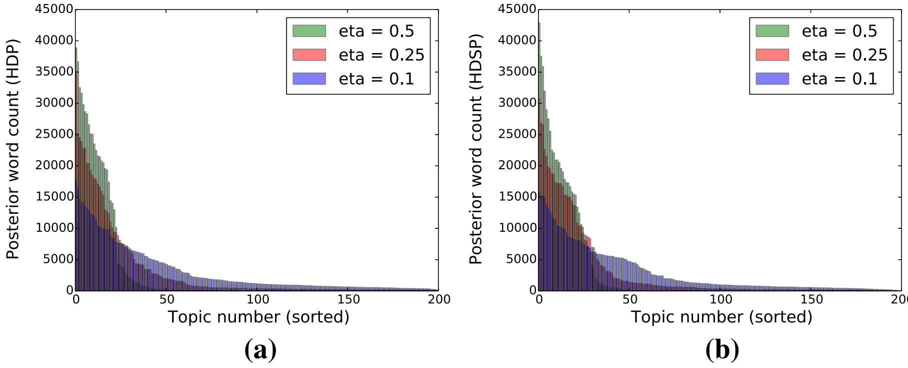


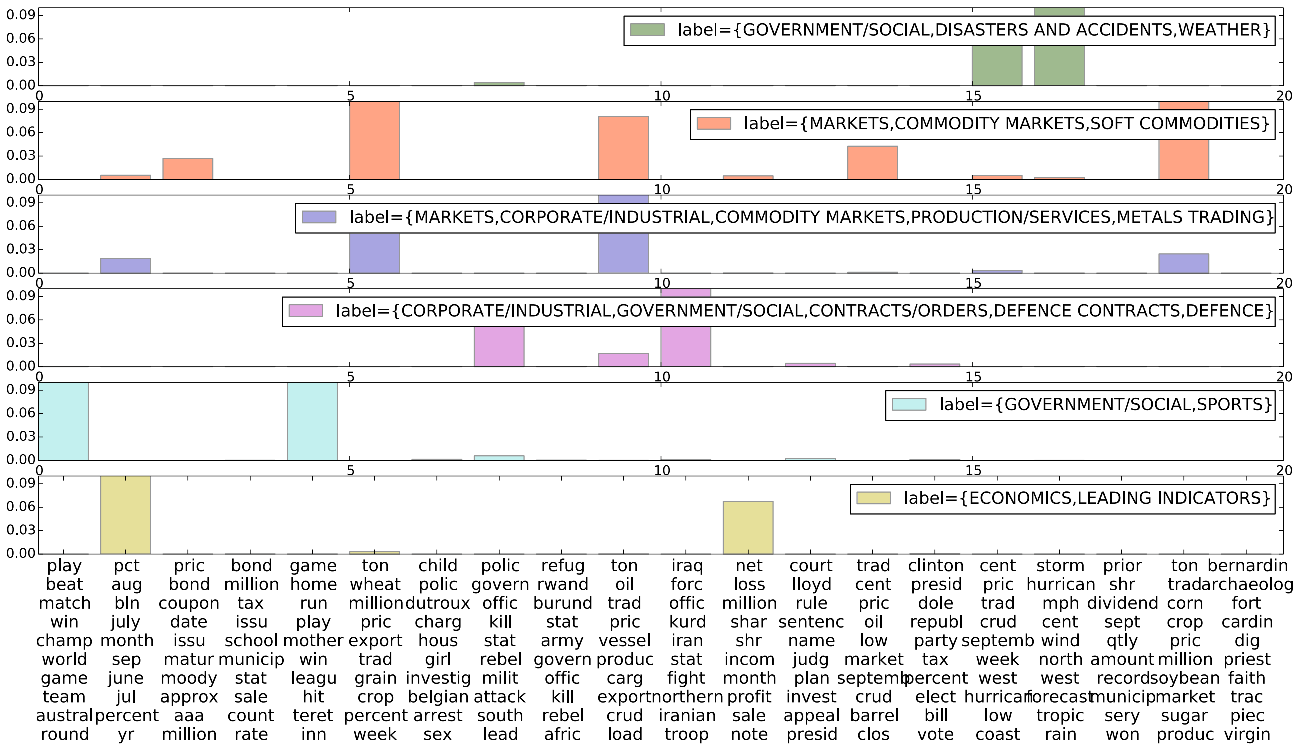
Fig. 11 Posterior number of words. We set the truncation level to 200, but only a few topics are used during inference. **a** HDP, **b** HDSP

Appendix 2: Posterior word count

Like the HDP and other nonparametric topic models, our model also uses only a few topics even though we set the truncation level to 200. Figure 11 shows the posterior word count for the different values of the Dirichlet topic parameter η. As the result indicates our model uses 50–100 topics. The HDSP tends to use more topics than the HDP.

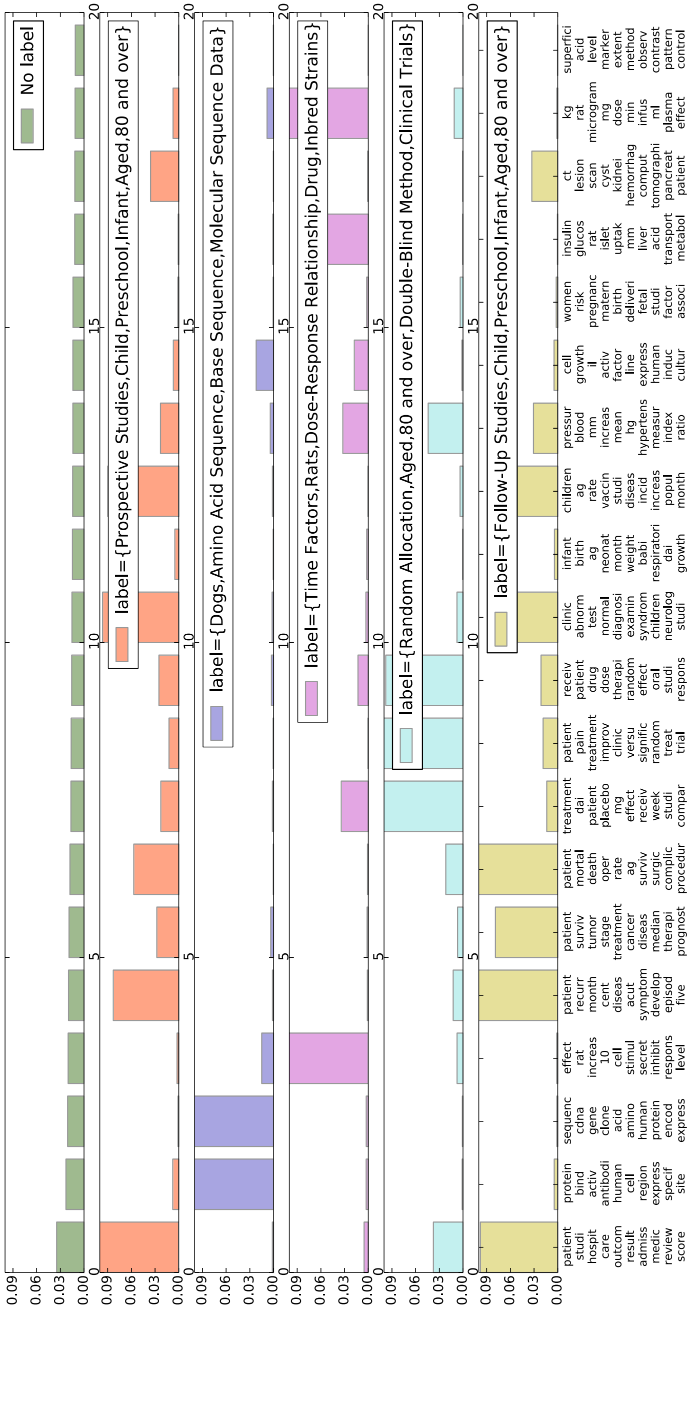
Appendix 3: Visualize the correlation between labels and topics

We visualise the inferred correlation between labels and topics by computing the expected topic distribution given a set of labels. Figures 12 and 13 show the expected topic distributions with respect to different sets of labels. When multiple labels are given, the model expects high probabilities for the topics that are similar to all given labels. For example, when 'Market' and 'Sports' labels are given, the model expects high probabilities on sports related topics and relatively high probability on 'Market' related topics based on the weights between topics and two labels.



(a)

Fig. 12 Expected topic distributions given labels from RCV and OHSUMED. Topics are sorted by their posterior word counts, and the top 20 topics are displayed with the top 10 words (stemmed). From *top to bottom*, we compute an expected topic distribution given a randomly selected set of labels. **a** RCV, **b** OHSUMED



(b)

Fig. 12 continued

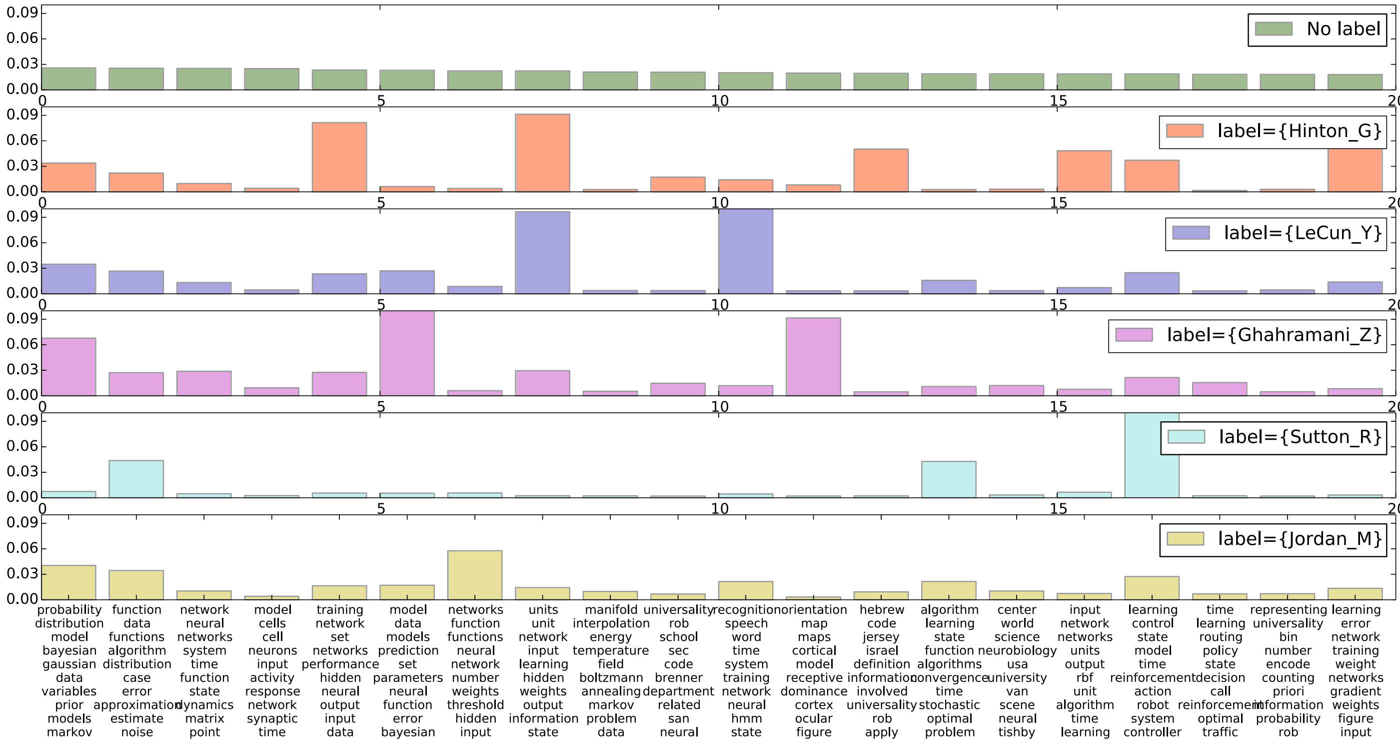


Fig. 13 Expected topic distributions given labels from NIPS. Topics are sorted by their posterior word counts, and the top 20 topics are displayed with the top 10 words. From *top to bottom*, we compute an expected topic distribution given a set of representative labels

References

Ahmed, A., & Xing, E. P. (2010). Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the 26th conference on uncertainty in artificial intelligence (UAI)* (pp. 20–29).

Ahn, S., Shahbaba, B., & Welling, M. (2014). Distributed stochastic gradient mcmc. In *Proceedings of the 31th international conference on machine learning (ICML)*.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 1). New York: springer.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for dirichlet process mixtures. *Bayesian Analysis*, *1*(1), 121–144.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics,* 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research,* 993–1022.

Chang, J., & Blei, D. M. (2009). Relational topic models for document networks. In *International conference on artificial intelligence and statistics* (pp. 81–88).

Duan, J. A., Guindani, M., & Gelfand, A. E. (2007). Generalized spatial dirichlet process models. *Biometrika*, *94*(4), 809–825.

Dunson, D. B., & Park, J. H. (2008). Kernel stick-breaking processes. *Biometrika*, *95*(2), 307–323.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*, 577–588.

Gelfand, A. E., Kottas, A., & MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association*, *100*(471), 1021–1035.

Griffin, J. E., & Steel, M. J. (2006). Order-based dependent dirichlet processes. *Journal of the American statistical Association*, *101*(473), 179–194.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, *14*(1), 1303–1347.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233.

Kim, D., Kim, S., & Oh, A. (2012). Dirichlet process with mixed random measures: A nonparametric topic model for labeled data. In *Proceedings of the 29th international conference on machine learning (ICML)*.

Kim, D. I., & Sudderth, E. B. (2011). The doubly correlated nonparametric topic model. *Advances in Neural Information Processing Systems,* 1980–1988.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *International conference on learning representations (ICLR)*.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Liang, P., Petrov, S., Jordan, M. I., & Klein, D. (2007). The infinite pcfg using hierarchical dirichlet processes. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 688–697).

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science* (pp. 50–55).

Mcauliffe, J. D., & Blei, D. M. (2007). Supervised topic models. *Advances in Neural Information Processing Systems*.

Mimno, D., & McCallum, A. (2012). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. arXiv:1206.3278.

Mimno, D., Li, W., & McCallum, A. (2007). Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on machine learning (ICML)* (pp. 633–640). ACM.

Paisley, J., Wang, C., & Blei, D. M. (2012). The discrete infinite logistic normal distribution. *Bayesian Analysis*, *7*(4), 997–1034.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (Vol. 1, pp. 248–256). Association for Computational Linguistics.

Ramage, D., Manning, C. D., & Dumais, S. (2011). Partially labeled topic models for interpretable text mining. *Proceedings of the 17th ACM international conference on knowledge discovery and data mining (KDD)* (pp. 457–465). New York, NY.

Ranganath, R., Gerrish, S., & Blei, D. (2014). Black box variational inference. In *Proceedings of the seventeenth international conference on artificial intelligence and statistics (AISTATS)* (pp. 814–822).

Rao, V., & Teh, Y. W. (2009). Spatial normalized gamma processes. *Advances in Neural Information Processing Systems,* 1554–1562.
Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian processes for machine learning (adaptive computation and machine learning)*. Cambridge: The MIT Press.
Ren, L., Du, L., Carin, L., & Dunson, D. (2011). Logistic stick-breaking process. *The Journal of Machine Learning Research*, *12*, 203–239.
Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *UAI*.
Sethuraman, J. (1991). A constructive definition of dirichlet priors. *Statistica Sinica*, *4*, 639–650.
Srebro, N., & Roweis, S. (2005). Time-varying topic models using dependent dirichlet processes. UTML, TR# 2005 3.
Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. Journal of the American Statistical Association.
Teh, Y. W., Kurihara, K., & Welling, M. (2008). Collapsed variational inference for HDP. NIPS 20.
Tran, D., Ranganath, R., & Blei, D. M. (2016). Variational gaussian process. In *International conference on learning representations (ICLR)*.
Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, *1*(1–2), 1–305.
Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th international conference on machine learning*.
Zhu, J., Ahmed, A., & Xing, E. P. (2009). Medlda: Maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1257–1264). ACM.